Exercise4

```
#Install and load the arrow package
#install.packages("arrow")
#install.packages("gender")
#install.packages("devtools")
#devtools::install_github("ropensci/genderdata", type = "source")
#install_genderdata_package()
```

```
library(genderdata)
```

```
library(gender)
```

```
## Warning: package 'gender' was built under R version 4.3.3
```

```
library(arrow)
```

```
## Warning: package 'arrow' was built under R version 4.3.3
```

```
##
## Attaching package: 'arrow'
```

```
## The following object is masked from 'package:utils':
##
##     timestamp
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(wru)
```

```
## Warning: package 'wru' was built under R version 4.3.3
```

```
##
## Please cite as:
##
## Khanna K, Bertelsen B, Olivella S, Rosenman E, Rossell Hayes A, Imai K
## (2024). _wru: Who are You? Bayesian Prediction of Racial Category Using
## Surname, First Name, Middle Name, and Geolocation_. R package version
## 3.0.1, <https://CRAN.R-project.org/package=wru>.
##
## Note that wru 2.0.0 uses 2020 census data by default.
## Use the argument ‘year = "2010"‘, to replicate analyses produced with earlier package versions.
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.3.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:arrow':
##
##     duration
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(igraph)
```

```
## Warning: package 'igraph' was built under R version 4.3.3
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
```

```
## The following object is masked from 'package:tidyr':
##
##     crossing
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```r
# set option to view all columns
options(dplyr.width = Inf)
```

```r
# Data ingestion from stored files in Parquet and CSV formats for analysis.
parquet_file <- "C:/Users/pc/Downloads/Project_Network_Analysis/app_data_sample.parquet"
applications  <- read_parquet(parquet_file)
```

```r
# Read CSV file
edge_link <- "C:/Users/pc/Downloads/Project_Network_Analysis/edges_sample.csv"
edges <- read.csv(edge_link)
```

```r
# Quick structure overview of the loaded 'applications' dataframe.
str(applications)
```

```
## tibble [2,018,477 x 16] (S3: tbl_df/tbl/data.frame)
##  $ application_number  : chr [1:2018477] "08284457" "08413193" "08531853" "08637752" ...
##  $ filing_date         : Date[1:2018477], format: "2000-01-26" "2000-10-11" ...
##  $ examiner_name_last  : chr [1:2018477] "HOWARD" "YILDIRIM" "HAMILTON" "MOSHER" ...
##  $ examiner_name_first : chr [1:2018477] "JACQUELINE" "BEKIR" "CYNTHIA" "MARY" ...
##  $ examiner_name_middle: chr [1:2018477] "V" "L" NA NA ...
##  $ examiner_id         : num [1:2018477] 96082 87678 63213 73788 77294 ...
##  $ examiner_art_unit   : num [1:2018477] 1764 1764 1752 1648 1762 ...
##  $ uspc_class          : chr [1:2018477] "508" "208" "430" "530" ...
##  $ uspc_subclass       : chr [1:2018477] "273000" "179000" "271100" "388300" ...
##  $ patent_number       : chr [1:2018477] "6521570" "6440298" "5607816" "6927281" ...
##  $ patent_issue_date   : Date[1:2018477], format: "2003-02-18" "2002-08-27" ...
##  $ abandon_date        : Date[1:2018477], format: NA NA ...
##  $ disposal_type       : chr [1:2018477] "ISS" "ISS" "ISS" "ISS" ...
##  $ appl_status_code    : num [1:2018477] 150 250 250 250 161 150 135 161 161 250 ...
##  $ appl_status_date    : chr [1:2018477] "30jan2003 00:00:00" "27sep2010 00:00:00" "30mar2009 00:00:
##  $ tc                  : num [1:2018477] 1700 1700 1700 1600 1700 1700 1600 1600 1600 1700 ...
```

Identifying Examiner Genders

We aim to infer the gender of each examiner from their first name, listed in the examiner_name_first column, utilizing the gender library based on an adapted version of an example they provided.

The applications database houses over 2 million entries. This is reflective of the numerous entries per examiner, correlating with the total applications they have assessed in the given period. Our initial step is to compile a distinct list of examiner names in examiner_names. Subsequently, we'll estimate the gender for each unique name and reintegrate this data into the original dataset. To begin, we focus on extracting unique names to avoid redundancy:

```r
# Extract unique examiner first names.
examiner_names <- applications %>%
  distinct(examiner_name_first)

# Associate names with gender and simplify the resulting data.
```

```
examiner_names_gender <- examiner_names %>%
  do(results = gender(.$examiner_name_first, method = "ssa")) %>%
  unnest(cols = c(results), keep_empty = TRUE) %>%
  select(
    examiner_name_first = name,
    gender,
    proportion_female
  )

# remove extra colums from the gender table
examiner_names_gender <- examiner_names_gender %>%
  select(examiner_name_first, gender)

# # Merge the gender information back into the main dataset.
applications <- applications %>%
  left_join(examiner_names_gender, by = "examiner_name_first")

# Cleanup of intermediate variables to free memory.
rm(examiner_names)
rm(examiner_names_gender)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  4486571 239.7    8184904 437.2  4506840 240.7
## Vcells 49528483 377.9   93023166 709.8 79844581 609.2
```

Guess the examiner's race

Similar process as gender identification, focusing on last names to predict race using the 'predict_race' function from the 'wru' package.

```
examiner_surnames <- applications %>%
  select(surname = examiner_name_last) %>%
  distinct()

examiner_race <- predict_race(voter.file = examiner_surnames, surname.only = T) %>%
  as_tibble()
```

```
## Predicting race for 2020

## Warning: Unknown or uninitialised column: 'state'.

## Proceeding with last name predictions...

## i All local files already up-to-date!

## 701 (18.4%) individuals' last names were not matched.
```

```r
examiner_race <- examiner_race %>%
  mutate(max_race_p = pmax(pred.asi, pred.bla, pred.his, pred.oth, pred.whi)) %>%
  mutate(race = case_when(
    max_race_p == pred.asi ~ "Asian",
    max_race_p == pred.bla ~ "black",
    max_race_p == pred.his ~ "Hispanic",
    max_race_p == pred.oth ~ "other",
    max_race_p == pred.whi ~ "white",
    TRUE ~ NA_character_
  ))

# removing extra columns
examiner_race <- examiner_race %>%
  select(surname,race)

applications <- applications %>%
  left_join(examiner_race, by = c("examiner_name_last" = "surname"))

rm(examiner_race)
rm(examiner_surnames)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  4596116 245.5    8184904 437.2  6109035 326.3
## Vcells 51746989 394.8   93023166 709.8 91024749 694.5
```

Data preparation: Converting date columns to Date type and calculating processing time for applications.

```r
applications <- applications %>%
  mutate(
    filing_date = as.Date(filing_date),
    patent_issue_date = as.Date(patent_issue_date),
    abandon_date = as.Date(abandon_date),
    final_decision_date = coalesce(patent_issue_date, abandon_date),
    app_proc_time = as.numeric(final_decision_date - filing_date),
    # Replace negative app_proc_time with NA
    app_proc_time = ifelse(app_proc_time < 0, NA, app_proc_time)
  )
```

Load additional libraries for network analysis and graph-based visualization.

```r
library(dplyr)
library(tidygraph)
```

```
##
## Attaching package: 'tidygraph'

## The following object is masked from 'package:igraph':
##
##     groups

## The following object is masked from 'package:stats':
##
##     filter
```

```
library(ggraph)
```

```
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
# Preparing edge data for network analysis, including handling missing values and data type conversions

edges <- edges %>%
  mutate(
    from = as.character(ego_examiner_id),
    to = as.character(alter_examiner_id)
  ) %>%
  mutate(
    from = ifelse(is.nan(as.numeric(from)), NA, from),
    to = ifelse(is.nan(as.numeric(to)), NA, to)
  ) %>%
  drop_na()

# Relocating and renaming examiner ID for consistency in the applications dataset, ensuring data cleanl

applications <- applications %>%
  relocate(examiner_id, .before = application_number) %>%
  mutate(examiner_id = as.character(examiner_id)) %>%
  drop_na(examiner_id) %>%
  rename(name = examiner_id)

# Constructing a graph object for network analysis, ensuring nodes are unique and data is consistent.

graph <- tbl_graph(
  edges = (edges %>% relocate(from, to)),
  directed = TRUE
)

applications <- applications %>%
  mutate(name = as.character(name)) %>%
  distinct(name, .keep_all = TRUE)

graph <- graph %>%
  activate(nodes) %>%
  inner_join(
```

```
    (applications %>% distinct(name, .keep_all = TRUE)),
    by = "name"
  )
```

Calculating network centrality measures for nodes, aiding in the analysis of examiner influence and connectivity.

```
graph %>%
  activate(nodes) %>%
  mutate(
    degree = centrality_degree(),
    betweenness = centrality_betweenness(),
    closeness = centrality_closeness()
  ) %>%
  select(name, degree, betweenness, closeness) %>%
  arrange(-degree)
```

```
## # A tbl_graph: 2504 nodes and 17809 edges
## #
## # A directed multigraph with 130 components
## #
## # A tibble: 2,504 x 4
##   name  degree betweenness closeness
##   <chr>  <dbl>       <dbl>     <dbl>
## 1 83670    198          0   0.000403
## 2 97910    176        132.  0.00787
## 3 73920    174          0   0.00971
## 4 67226    122        876.  0.00746
## 5 80730    120          0   0.000286
## 6 75615    117          0   0.000457
## # i 2,498 more rows
## #
## # A tibble: 17,809 x 6
##    from    to application_number advice_date ego_examiner_id alter_examiner_id
##   <int> <int>              <int> <chr>                <int>             <int>
## 1   158  1462            9402488 2008-11-17           84356             66266
## 2   158  1463            9402488 2008-11-17           84356             63519
## 3   158  1464            9402488 2008-11-17           84356             98531
## # i 17,806 more rows
```

Integrating centrality measures back into the applications dataframe for comprehensive analysis.

```
node_data <- graph %>%
  activate(nodes) %>%
  mutate(
```

```
    degree = centrality_degree(),
    betweenness = centrality_betweenness(),
    closeness = centrality_closeness()
  ) %>%
  select(name, degree, betweenness, closeness) %>%
  as_tibble() # Convert to a tibble/data frame for joining

# Joining the centrality measures back to the applications dataframe
applications <- applications %>%
  left_join(node_data, by = c("name" = "name"))

# rename name to examiner_id
applications <- applications %>%
  rename(examiner_id = name)

head(applications,5)
```

```
## # A tibble: 5 x 23
##   examiner_id application_number filing_date examiner_name_last
##   <chr>       <chr>              <date>      <chr>
## 1 96082       08284457           2000-01-26  HOWARD
## 2 87678       08413193           2000-10-11  YILDIRIM
## 3 63213       08531853           2000-05-17  HAMILTON
## 4 73788       08637752           2001-07-20  MOSHER
## 5 77294       08682726           2000-04-10  BARR
##   examiner_name_first examiner_name_middle examiner_art_unit uspc_class
##   <chr>               <chr>                             <dbl> <chr>
## 1 JACQUELINE          V                                  1764 508
## 2 BEKIR               L                                  1764 208
## 3 CYNTHIA             <NA>                               1752 430
## 4 MARY                <NA>                               1648 530
## 5 MICHAEL             E                                  1762 427
##   uspc_subclass patent_number patent_issue_date abandon_date disposal_type
##   <chr>         <chr>         <date>            <date>       <chr>
## 1 273000        6521570       2003-02-18        NA           ISS
## 2 179000        6440298       2002-08-27        NA           ISS
## 3 271100        5607816       1997-03-04        NA           ISS
## 4 388300        6927281       2005-08-09        NA           ISS
## 5 430100        <NA>          NA                2000-12-27   ABN
##   appl_status_code appl_status_date      tc gender race  final_decision_date
##              <dbl> <chr>              <dbl> <chr>  <chr> <date>
## 1              150 30jan2003 00:00:00  1700 female white 2003-02-18
## 2              250 27sep2010 00:00:00  1700 <NA>   white 2002-08-27
## 3              250 30mar2009 00:00:00  1700 female white 1997-03-04
## 4              250 07sep2009 00:00:00  1600 female white 2005-08-09
## 5              161 19apr2001 00:00:00  1700 male   white 2000-12-27
##   app_proc_time degree betweenness closeness
##           <dbl>  <dbl>       <dbl>     <dbl>
## 1          1119     NA          NA      NA
## 2           685     NA          NA      NA
## 3            NA      0           0     NaN
## 4          1481      2           0       0.5
## 5           261      0           0     NaN
```

```r
#null values in applications data each column
sapply(applications, function(x) sum(is.na(x)))
```

```
##          examiner_id   application_number          filing_date
##                    0                    0                    0
##   examiner_name_last  examiner_name_first examiner_name_middle
##                    0                    0                 1370
##    examiner_art_unit           uspc_class        uspc_subclass
##                    0                    0                    0
##        patent_number    patent_issue_date         abandon_date
##                 2606                 2605                 3399
##        disposal_type     appl_status_code     appl_status_date
##                    0                    1                    1
##                   tc               gender                 race
##                    0                  799                    0
##  final_decision_date        app_proc_time               degree
##                  356                  357                 3144
##          betweenness            closeness
##                 3144                 4211
```

```r
# total rows in applications data
nrow(applications)
```

```
## [1] 5648
```

```r
# Dropping rows with NA in regression columns
applications <- applications %>%
  drop_na(app_proc_time, degree, gender, examiner_art_unit, uspc_class,disposal_type,race)
```

Model preparation: Transforming selected variables into categorical factors for regression analysis.

```r
applications <- applications %>%
  mutate(
    examiner_art_unit = as.factor(examiner_art_unit),
    uspc_class = as.factor(uspc_class),
    gender = as.factor(gender),
    race = as.factor(race),
    disposal_type = as.factor(disposal_type)
  )
```

I wanted to use examiner_art_unit, uspc_class as categorical variable but considering ther are too many they are not added as features

disposal_type categorical varaible is used because it tellls about the status of appplication“ISS” (issued), “ABN” (abandoned), “PEND’ (PENDING). There must be a differnce in processing times for each of the category

Race is used as well to understand affect of race in processing times

```r
#Model 1: Examining the influence of degree centrality along with categorical variables on application p

model_degree <- lm(app_proc_time ~ degree +race +disposal_type , data = applications)
summary(model_degree)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree + race + disposal_type, data = applications)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2464.0  -808.2  -240.2   678.9  4275.1
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1869.99      50.59  36.963  < 2e-16 ***
## degree              10.29       1.49   6.908 6.51e-12 ***
## raceblack           96.09     135.31   0.710   0.4777
## raceHispanic        26.14     146.64   0.178   0.8586
## raceother         -542.76     751.52  -0.722   0.4702
## racewhite         -126.15      50.89  -2.479   0.0133 *
## disposal_typeISS    92.74      47.03   1.972   0.0488 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1061 on 2088 degrees of freedom
## Multiple R-squared:  0.02804,    Adjusted R-squared:  0.02525
## F-statistic: 10.04 on 6 and 2088 DF,  p-value: 5.821e-11
```

Model 1: Degree Centrality with Categorical Variables Model Formula: `app_proc_time ~ degree + race + disposal_type`

The results from Model 1 provide several insights into the relationship between the application processing time and the examined variables, including degree centrality, race, and disposal type of the patent application. Here's a breakdown of the key points:

Degree Centrality: The coefficient for degree centrality is positive (Estimate = 10.29) and statistically significant ($p < 0.001$). This suggests that for each unit increase in degree centrality, the processing time increases by approximately 10.29 days, holding other variables constant. This could indicate that examiners who are more central to the network—perhaps because they handle a greater volume of applications or are involved in more complex cases—tend to have longer processing times.

Race: The coefficients for the race categories show mixed results. Notably, applications processed by examiners identified as "white" have processing times that are, on average, 126.15 days shorter than those processed by examiners of the baseline race category (which is not specified here but could be implied as the omitted category), and this result is statistically significant ($p = 0.0133$). The effects for other race categories ("black," "Hispanic," "other") are not statistically significant, indicating that, compared to the baseline category, these races do not have a significantly different processing time when controlling for other factors.

Disposal Type: The coefficient for disposal type "ISS" (issued patents) is positive (Estimate = 92.74) and statistically significant ($p = 0.0488$), suggesting that applications that are eventually issued take, on average, 92.74 days longer to process than those that are not issued, perhaps reflecting the additional scrutiny and requirements involved in issuing a patent.

Model Fit: The model's R-squared value is 0.02804, indicating that approximately 2.8% of the variance in processing times can be explained by the model's variables. While statistically significant, this suggests that the majority of the variation in processing times is due to factors not included in the model.

Overall Significance: The F-statistic (10.04) and its associated p-value (5.821e-11) indicate that the model is statistically significant, meaning that there is a relationship between the predictor variables and processing time. However, given the low R-squared value, the model's explanatory power is limited.

In summary, Model 1 highlights the significance of an examiner's network position and the racial categorization of the examiner on the processing times of patent applications, with specific attention to the differential impact on applications that are issued. Despite its statistical significance, the model explains a relatively small portion of the variability in processing times, suggesting that additional factors not captured by the model may play a substantial role in determining processing outcomes.

```
#Model 2: Betweenness Centrality with Categorical Variables

model_betweenness <- lm(app_proc_time ~ betweenness +race +disposal_type, data = applications)
summary(model_betweenness)
```

```
##
## Call:
## lm(formula = app_proc_time ~ betweenness + race + disposal_type,
##     data = applications)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1761.5  -797.8  -232.1   687.1  4219.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.943e+03  5.008e+01  38.798   <2e-16 ***
## betweenness       8.116e-03  8.704e-03   0.932   0.3512
## raceblack         8.642e+01  1.368e+02   0.632   0.5278
## raceHispanic      4.138e+01  1.483e+02   0.279   0.7803
## raceother        -5.457e+02  7.599e+02  -0.718   0.4728
## racewhite        -1.268e+02  5.152e+01  -2.461   0.0139 *
## disposal_typeISS  8.613e+01  4.755e+01   1.811   0.0702 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1073 on 2088 degrees of freedom
## Multiple R-squared:  0.006245,   Adjusted R-squared:  0.003389
## F-statistic: 2.187 on 6 and 2088 DF,  p-value: 0.04158
```

Model 2: Betweenness Centrality with Categorical Variables Model Formula: `app_proc_time ~ betweenness + race + disposal_type`

Model 2 evaluates the impact of betweenness centrality, race, and disposal type on patent application processing times. The findings are as follows:

Betweenness Centrality: The coefficient for betweenness centrality is not statistically significant (Estimate = 0.0081, p = 0.3512), suggesting it has a negligible impact on processing times. This implies that an examiner's role as a network bridge does not significantly affect the speed of processing patent applications.

Race: The coefficient for examiners identified as "white" shows a significant reduction in processing times (Estimate = -126.8, p = 0.0139), similar to Model 1, indicating that racial categorization influences processing durations. Other racial categories did not show a statistically significant difference.

Disposal Type: Applications that are issued (disposal_typeISS) show a tendency towards longer processing times (Estimate = 86.13, p = 0.0702), although this result is marginally significant, hinting at the extensive review required for issuance.

Model Fit: The R-squared value is notably low at 0.006245, indicating that only about 0.62% of the variance in processing times is explained by the model. This highlights the presence of other influential factors not captured by this model.

Overall Significance: Despite the low explanatory power, the model is statistically significant (p-value = 0.04158), suggesting that the variables included do have an effect on processing times, albeit a small one.

In essence, Model 2 underscores the limited role of betweenness centrality in affecting patent processing times, reaffirms the impact of racial categorization, and suggests a nuanced influence of disposal type on processing durations. However, the model's low R-squared value points to a significant portion of the variability in processing times being driven by factors outside the model's scope.

```
#Model 3: Degree Centrality with Gender Interaction and Categorical Variables

model_degree_gender <- lm(app_proc_time ~ degree * gender + +race +disposal_type, data = applications)
summary(model_degree_gender)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree * gender + +race + disposal_type,
##     data = applications)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2336.0  -796.3  -236.9   667.6  4388.8
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1762.076     63.567  27.720  < 2e-16 ***
## degree              12.533      2.736   4.581  4.9e-06 ***
## gendermale         157.424     56.247   2.799  0.00518 **
## raceblack          107.099    135.196   0.792  0.42835
## raceHispanic        25.608    146.580   0.175  0.86133
## raceother         -584.969    750.637  -0.779  0.43589
## racewhite         -132.025     50.887  -2.594  0.00954 **
## disposal_typeISS    90.661     46.977   1.930  0.05375 .
## degree:gendermale   -3.204      3.262  -0.982  0.32606
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1060 on 2086 degrees of freedom
## Multiple R-squared:  0.03169,    Adjusted R-squared:  0.02798
## F-statistic: 8.534 on 8 and 2086 DF,  p-value: 1.712e-11
```

Model 3: Degree Centrality with Gender Interaction Model Formula: `app_proc_time ~ degree * gender + race + disposal_type`

Model 3 explores how degree centrality, gender, race, and disposal type impact patent application processing times, incorporating an interaction term between degree centrality and gender. The results indicate:

Degree Centrality: Each unit increase in degree centrality increases processing time by approximately 12.53 days ($p < 0.00001$), indicating that more central examiners experience longer processing times, likely due to handling more or complex applications.

Gender: Male examiners on average have processing times that are 157.42 days longer than their female counterparts ($p = 0.00518$), highlighting a significant gender disparity in processing times.

Race: Similar to previous models, the race category "white" shows a significant reduction in processing times (-132.02 days, $p = 0.00954$), reaffirming racial differences in processing speeds. Other racial categories do not show significant differences.

Disposal Type: The coefficient for issued patents (disposal_typeISS) is positive (90.66) and approaches statistical significance ($p = 0.05375$), suggesting that patents that are issued tend to have longer processing times, albeit less conclusively than in prior models.

Degree and Gender Interaction: The interaction term between degree centrality and gender (male) is not significant ($p = 0.32606$), suggesting that the impact of degree centrality on processing times does not differ significantly between male and female examiners.

Model Fit: The model explains 3.17% of the variance in processing times (Multiple R-squared = 0.03169), a slight improvement over the previous models but still indicating a large portion of variance is unexplained by the model's variables.

Overall Significance: The model is statistically significant (p-value = 1.712e-11), indicating a reliable relationship between the predictors and processing times, despite the low explanatory power.

In summary, Model 3 underscores the influence of examiner centrality and gender on processing times, with significant findings for gender differences and the impact of being a "white" examiner. The interaction between degree and gender does not significantly affect processing times, suggesting the primary effects of centrality and gender operate independently of each other. Despite its contributions, the model leaves much of the variance in processing times unexplained, pointing to the complexity of factors influencing patent processing.

```
#Model 4: Betweenness Centrality with Gender Interaction and Categorical Variables
model_betweenness_gender <- lm(app_proc_time ~ betweenness * gender +race +disposal_type, data = applica
summary(model_betweenness_gender)
```

```
##
## Call:
## lm(formula = app_proc_time ~ betweenness * gender + race + disposal_type,
##     data = applications)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1740.8  -807.7  -242.5   685.1  4312.9
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1854.34166   61.40286  30.200   <2e-16 ***
## betweenness              -0.02409    0.02883  -0.835   0.4035
## gendermale              128.72866   52.47789   2.453   0.0142 *
## raceblack                96.87994  136.71236   0.709   0.4786
## raceHispanic             34.89838  148.13060   0.236   0.8138
## raceother              -585.87196  759.00019  -0.772   0.4403
## racewhite              -131.82748   51.52873  -2.558   0.0106 *
```

```
## disposal_typeISS         86.60328   47.50330   1.823   0.0684 .
## betweenness:gendermale    0.03452    0.03025   1.141   0.2539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1071 on 2086 degrees of freedom
## Multiple R-squared:  0.01004,    Adjusted R-squared:  0.006245
## F-statistic: 2.645 on 8 and 2086 DF,  p-value: 0.006933
```

‘‘‘

Model 4: Betweenness Centrality with Gender Interaction Model 4 assesses the impact of betweenness centrality and its interaction with gender, alongside race and disposal type, on patent application processing times. The findings indicate:

Betweenness Centrality: The coefficient for betweenness centrality (-0.02409) is not significant (p = 0.4035), suggesting that betweenness centrality alone does not have a clear impact on processing times. This implies that an examiner's position as a network bridge does not significantly influence how quickly they process applications.

Gender: The coefficient for male examiners (128.73) is significant (p = 0.0142), indicating that male examiners, on average, have longer processing times than their female counterparts. This highlights a gender disparity in processing times.

Race: Consistent with previous models, the race category "white" shows a significant reduction in processing times (-131.83, p = 0.0106), reiterating that racial differences exist in processing speeds. Other racial categories do not show significant differences.

Disposal Type: The coefficient for issued patents (disposal_typeISS) is positive (86.60) and approaches statistical significance (p = 0.0684), suggesting a trend where issued patents may require longer processing times, although this result is less conclusive.

Betweenness and Gender Interaction: The interaction between betweenness centrality and gender (male) is not significant (p = 0.2539), indicating that the effect of betweenness centrality on processing times does not significantly differ between male and female examiners.

Model Fit: The R-squared value is 0.01004, meaning that the model explains only about 1% of the variance in processing times. This indicates that a vast majority of the variance is due to factors not included in the model.

Overall Significance: Despite the low explanatory power, the model is statistically significant (p-value = 0.006933), suggesting there are relationships between the predictors and processing times, albeit small.

In essence, Model 4 underscores a lack of significant impact from betweenness centrality on processing times and confirms the influence of gender and racial categorization. The interaction term between betweenness and gender does not significantly influence processing times, suggesting that the primary effects of gender and betweenness operate independently. The low R-squared value highlights the complexity of factors influencing patent processing times, indicating that many influencing variables are not captured by the model.

Explaining Regression Results and Implications for the USPTO

Conclusion and Implications for the USPTO

The series of linear regression models were designed to examine how the centrality of examiners within the USPTO network and other demographic factors affect patent application processing times. Key findings indicate that higher degree centrality marginally increases processing times, while betweenness centrality does not demonstrate a significant impact. This suggests that examiners who occupy more central roles

in the network, potentially handling a larger volume or more complex applications, may experience longer processing times.

Implications of Centrality on Operational Efficiency

**1. Strategic Workload Management: The observed correlation between degree centrality and longer processing times signals a need for strategic workload management. Examiners at the network's core may benefit from targeted support to manage their heavier or more complex caseloads efficiently. Implementing strategies such as workload redistribution or increased support could help maintain or even improve the quality of patent examination without extending processing times.

Examination of Gender Interaction

Gender-Based Processing Time Differences: The analysis did not find significant interaction effects between gender and centrality on processing times, yet it did reveal that male examiners generally have longer processing times than their female counterparts. This distinction raises questions about underlying factors contributing to these differences and suggests the need for further investigation into the workflows and challenges faced by examiners of different genders

Implications for the USPTO

Equitable Process Optimization: The findings underscore the importance of considering both centrality in the examiner network and gender when devising strategies to optimize processing times. With the goal of ensuring equitable and efficient patent examination processes, the USPTO may need to adopt tailored approaches that account for the unique challenges and needs of examiners based on their network position and gender.

Addressing Racial Disparities: Similarly, the significant differences in processing times across racial categories suggest that racial dynamics within the examination process warrant closer attention. Understanding and addressing these disparities is crucial for fostering an equitable work environment and ensuring that patent examination standards remain consistent across diverse groups of examiners.

In conclusion, while the regression models highlight several factors influencing patent application processing times, they also reveal areas where the USPTO could focus its efforts to improve operational efficiency and equity. By acknowledging and addressing the nuanced effects of examiner centrality, gender, and race on processing times, the USPTO can take meaningful steps toward optimizing its patent examination processes and upholding its commitment to fairness and quality in intellectual property examination.